

Compare Different Methods for Extensibility of Measurement Results of Point Samples of the Soil Protection Information and Monitoring System

László Várallyai (University of Debrecen, Faculty of Agricultural Economics and Rural Development)

Béla Kovács (University of Debrecen Faculty of Agricultural Science)

Miklós Herdon (University of Debrecen, Faculty of Agricultural Economics and Rural Development)

Abstract. The Hungarian Soil Information Monitoring System (SIM) covers the whole country and provides opportunity to create similar information systems for the natural resources (atmosphere, supply of water, flora biological resources etc). The aim of the SIM is to relate these databases.

The SIM territorial measuring grid consists of 1236 measuring points with 21 elements content in Hungary. These points are representatives. Distribution of the points by soil types represents the variety of soil types of the country. The SIM is the essential basis of rational agri-environmental management and at the same time is an integrated unit of the environmental diagnosis of soils.

Our aim is developing a statistical based information system from the data of the measured SIM points. We developed a method for estimating element content. To determine the concentration of the elements, need only the GPS co-ordinates of the place based on the number of nearest neighbouring points. This method does not calculate with spatial circumstances. The other possibility is using the kriging method (spatial interpolation) for estimating more precisely the element content. In this study these two methods are compared.

After building our statistical based information system we can develop an Internet-based service that makes it possible to reach the objectives through arranging the results of analyses into a database.

Based on available data, the developed Internet-based service makes it possible to estimate element content at a certain diagnostic point with some statistical errors; it can also be applied in analyzing effects of environmental pollution.

Keywords. Soil Information Monitoring System, kriging method, Internet-based service

Soil monitoring in Hungary

A large amount of soil information are available in Hungary as a result of long-term observations, various soil survey, analyses and mapping activities on national (1:500,000), regional (1:100,000), farm (1:10,000-1:25,000) and field level (1:5,000-1:10,000) during the last sixty years. Thematic soil maps are available for the whole country in the scale of 1:25,000 and for 70% of the agricultural area in the scale of 1:10,000.

There are at least three reasons why this rich soil database has been developed (Várallyai, 1993):

- the small size of the country (93,000 km²)
- the great importance of agriculture and soils in the national economy
- the historically "soil loving" character of the Hungarian people, and particularly the Hungarian farmers.

In the last years all existing soil data were organized into a computerized geographic soil information system, which consists of two main parts:

- The soil data bank, including 3 different types of information:
 - basic topographic information (geodetic data standards)
 - point information (measured, calculated, estimated or coded data on the various characteristics of soil profiles)
 - territorial information (1:25,000 scale thematic maps) and soil properties.
- The information system, including models on moisture and plant nutrient regimes of soils; susceptibility of soils to various soil degradation processes, etc.

The Soil Information and Monitoring System (SIM) is an independent subsystem of the integrated Environmental Information and Monitoring System (EIMS) (Soil Information Monitoring Professional Committee, 1995).

The Soil Information Monitoring System (SIM) covers the whole country and provides opportunity to create similar information systems for the natural resources (atmosphere, supply of water, flora and biological resources etc.). The aim is to relate these databases.

The SIM territorial measuring grid consists of 1236 measuring points. These points are representatives. Distribution of the points by soil types represents the variety of soil types of the country.

In consequence of the EU accession, the particular and objective survey of current soil condition is a very important question, which can be the beginning of the implementation of the modern agrarian environmental management program. This survey is not much use if the change of condition cannot be investigated continuously in systematic interval.

On the basis of the above mentioned point of view decision was born on creating the planned National Environmental Protection Information and Monitoring System. The first working subsystem was realized as the Soil Information Monitoring System (SIM) module.

Based on physiographical-soil-ecological units, 1236 "representative" observation points were selected (and exactly defined by geographical coordinates using GPS). There were 865 points on agricultural land, 183 points in forests and 189 points in environmentally threatened "hot spot" regions. The latter represented 12 different types of environmental hazards or particularly sensitive areas such as: degraded soils, ameliorated soils, drinking water supply areas, watersheds of important lakes and reservoirs, protected areas with particularly sensitive ecosystems, "hot spots" of industrial, agricultural, urban and transport pollution, military fields, areas affected by (surface) mining, waste (water) disposal affected spots.

Applied methods

Calculating method of distance, element content, relative error and confidence interval

After converting and rounding the measured data of the available TIM samples I carried out further calculations in order to determine the distance between the points, their element contents, their relative errors and confidence intervals. For the calculations I used the following connections:

The program calculates the distances by using the Pythagorean Theorem:

$$Z = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

where x_1, y_1 are the coordinates of the known point, x_2, y_2 are the coordinates of the unknown point.

The software determines the calculated figure of chemical element content in case of 10 nearest neighbouring points by using the following connection:

$$C_x = \frac{1/Z_1 * C_1 + 1/Z_2 * C_2 + \dots + 1/Z_{10} * C_{10}}{1/Z_1 + 1/Z_2 + \dots + 1/Z_{10}} \quad (2)$$

where Z_1, Z_2, \dots, Z_{10} are the distances of the known points correlated to the basis profile number,
 C_1, C_2, \dots, C_{10} measured chemical element content in the known points.

The program calculates the percentage deviation – relative error – on the basis of the following formula:

$$Deviation = ABS \left(100 * \left(\frac{Sz - M}{M} \right) \right) [\%] \quad (3)$$

where ABS is the absolute value function, „Sz” means the element content calculated by me, „M” means the measured element content.

The program invokes two Excel functions to calculate the confidence interval, one of them calculates the standard deviation (σ), and the other one calculates the confidence interval.

To calculate the confidence interval, it calculates the standard deviation first (σ):

$$\sigma = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}} \quad (4)$$

Formula of the confidence interval:

$$\bar{x} \pm 1,96 \left(\frac{\sigma}{\sqrt{n}} \right) \quad (5)$$

Kriging method

Kriging (Oliver and Webster, 1990) is an interpolation method that predicts unknown values of a random process. More precisely, a Kriging prediction is a weighted linear combination of all output values already observed. These weights depend on the distances between the input for which the output is to be predicted and the inputs already simulated. Kriging assumes that the closer the inputs are, the more positively correlated the outputs are. This assumption is modeled through the correlogram or the related variogram, discussed below (Alsamamra et.al., 2009).

In deterministic simulation, Kriging has an important advantage over regression analysis: Kriging is an exact interpolator; that is, predicted values at observed input values are exactly equal to the observed (simulated) output values. In random simulation, however, the observed output values are only estimates of the true values, so exact interpolation loses its intuitive appeal. Therefore regression uses OLS, which minimizes the residuals - squared and summed over all observations.

Effectively, geostatistical models directly estimate the variance-covariance matrix. Geostatistical techniques, such as Kriging rely upon an estimated variance-covariance matrix, followed by EGLS (estimated generalized least squares), and BLUP (best linear unbiased prediction). The simplest case assumes one can specify correctly the variance-covariance matrix as a function of distance only (Stein et.al, 2003). The most typical application involves the smooth interpolation of a surface at points other than those measured. Usually, the method assumes errors are 0 at the measured points but modifications allow for measurement errors at the measured points.

The first step in most geostatistical models is to estimate the variance-covariance matrix. While techniques exist to perform this directly, the most common technique involves the intermediate stage of computing the variogram (Bates et.al., 1996)

The empirical variogram begins with the pair-wise squared differences among all errors (or sometimes a sample of errors for large data sets) plotted against the distance between the elements of the pair. Positively correlated errors will show small pair-wise squared differences while almost independent errors will show larger differences. For positively correlated residuals, the empirical variogram tends to start off low at small distances and rise with distance up to a point where it levels off. From the variogram one can estimate the parameters of fitted variogram functions. If the process is stationary, equivalence exists between the fitted variogram functions and fitted covariance functions. Only a relatively small number of valid covariance functions exist which yield guaranteed positive definite estimated variance-covariance matrices (Bailey and Gatrell, 1995).

Results

The developed statistical method for estimating element content

I developed a method to estimate the element content. During the process I chose an optional point that will be considered as unknown in the process (in Fig. 1, I marked this point with "U"). I have the measured results for the unknown point, but I did not count with them, I proceeded as if I did not have those measured results.

Around the unknown points I determined the nearest ten points (this number can be changed in the program). I marked these nearest points in Figure 1 with "K".

The distance between the known points (K) and the point considered as unknown (U) can be calculated by using the Pythagorean Theorem (Eq. 1):

When I determined the distance between all the ten known and "unknown" points, I had ten distance data $z_1, z_2, z_3, \dots, z_{10}$, from which with linear estimation (by using Eq. 2), concentration (c_x) of the certain element can be estimated even in those places where there is no TIM diagnostic point.

If I calculate the concentration (c_x) of the certain element for the unknown point, I can compare it with the measured data by ICP-OES spectrometer for the certain element and I can determine the relative standard deviation (Eq. 3).

These steps have to be done first for the same point regarding the other elements which quantities exceed the demonstration line. After that, these steps have to be done again for each available TIM diagnostic point, considering that they are the unknown points in the experiment. After this, I will have the concentration value and its relative standard deviation for each measurable element above the demonstration line for each diagnostic point.

I applied two Excel functions to calculate the confidence interval, one of them (Eq. 4) calculates the standard deviation (σ) and the other calculates the confidence interval (Eq 5).

I visualized the received data on Excel worksheets in each case to make further data processing easier.

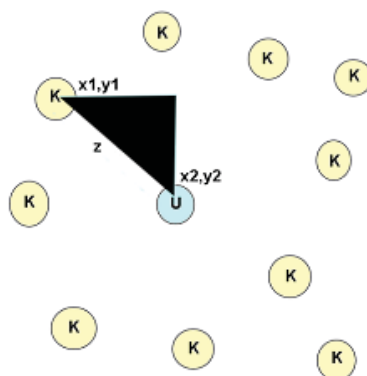


Figure 1. K = Known diagnostic points (x1,y1: GPS coordinates of the given place)
 U = diagnostic point to be determined (unknown)
 (x2,y2): GPS coordinates of the “unknown” place)

Building our statistical based information system has to determine the number of nearest neighbouring points to be considered in the case of certain elements. We have to determine the minimums of the average values of relative error for elements. In this point we have got the viewing neighbouring point number for all measured elements.

We discovered when studying different numbers (1, 2, 3, 4, 5, 6, 7, 10, 15, 20, 25 and 30 in order) of nearest neighbouring diagnostic points within the aggregation, the minimums of the relative error values are different. Elements can be ranged into three groups depending on how many nearest neighbouring diagnostic points were considered when we received the minimum value.

- 3 neighbouring diagnostical points: K, P, Sr, Ni, Cu, B, Co, Ti
- 5 neighbouring diagnostical points: Al, Fe, Ca, Mg, Mn, S, Ba, Cr, V, Pb, Y, Zn
- 10 neighbouring diagnostical points: Na

It means that in case of a sample originating from a genuinely unknown diagnostic point, for elements in the first column 3, for elements in the second column 5, while for Sodium 10 nearest neighboring points have to be considered in order to estimate the element content with the smallest error.

The kriging method for estimating element content

Calculation the concentration of the elements (by using Eq. 6) weighted average mathematical method (Korpás, 1996). The sum of the weighting factors must 0.

$$Z^*(x) = \sum_{i=1}^n a_i Z(x_i) \quad (6)$$

where $Z^*(x)$ the given concentration estimated value,
 a_i the weighted factor of the observed point

The weighting factors are inverse ratio to distance of the estimated locations (Bodrog, 2001).

$$a_i = \frac{\frac{1}{c + d_i^\omega}}{\sum_{i=1}^n \frac{1}{c + d_i^\omega}} \quad (7)$$

where d_i is the distance of the estimated location and the i -point,
 c is constant
 ω power (usually $1 < \omega < 3$)

We have to calculate the absolute error from the difference of the measured and estimated concentration value, and the relative error, which give the ratio of the measured and real concentration value (Young, 1986).

Building our statistical based information system has to determine the number of nearest neighbouring points to be considered in the case of certain elements.

We discovered that elements can be ranged only into two groups depending on how many nearest neighbouring diagnostic points were considered to kriging.

- 3 neighbouring diagnostical points: K, P, Sr
- 10 neighbouring diagnostical points: Al, B, Ba, Ca, Co, Cr, Cu, Fe, Mg, Mn, Na, Ni, Pb, S, Ti, V, Y, Zn

It means that using of kriging method in case of the most elements we need 10 nearest neighboring points have to be considered in order to estimate the element content with the smallest error. It is interesting, in case of K, P and Sr the results were the same (10 nearest neighboring points) in case of developed statistical method.

The Internet-based program

We developed a newer internet-based program that makes it possible to reach the objectives through arranging the results of analyses into a database and developing an authorisation system (Fig. 2.). When developing the newer software, my objective was to create a dynamic website for users which together with the GPS coordinates only of the certain point could provide the estimated element content for the measured chemical elements also in such points where no samples had been taken and thus measured data were not available. The relative error and the confidence interval must be provided for each element content value, since they orientate the users into which range the quantity of the certain chemical element at the given diagnostic point can fall.

We provided access to the database through internet-based technology. With the help of the database-based server-side script language (PHP) (Meloni, 2003) we planned and developed the user interface, through which the user can communicate with the program.

Figure 2. PHP form to evaluate the concentrations and reliability of the chosen elements based on the GPS-co-ordinates

Conclusion

On a statistical based developed system can be a suitable information system for the management determining the element concentrations in a well-defined precision. To determine the concentration of the elements, need only the GPS co-ordinates of the place.

Data can access by a suitable authority system in this information system. Through the authorization system, the authorization level of each user can easily be determined. In case of database administrator authorization, almost any kind of query can be made in connection with the database through the general query module.

Based on available data, the developed program makes it possible to estimate element content at a certain diagnostic point with some statistical errors; it can also be applied in analyzing effects of environmental pollution. Using by kriging method in case of the most elements we need 10 nearest neighboring points have to be considered in order to estimate the element content with the smallest error. It is interesting, in case of K, P and Sr the results were the same (10 nearest neighboring points) in case of the developed statistical method.

References

- Alsamamra, H., Ruiz-Arias, J.A., Pozo-Vázquez, D., Tovar-Pescador, J. 2009. Comparative study of ordinary and residual kriging techniques for mapping global solar radiation over southern Spain, *Agricultural and Forest meteorology*, 1-15
- Bailey, T., Gatrelli, A. 1995. Interactive Spatial Data Analysis, Harlow: Longman
- Bates, R.A., Buck, R.J., Riccomagno, E., Wynn, H.P. 1996. Experimental design and observation for large systems, *Journal of the Royal Statistical Society. Series (B58)* 77-94.
- Bodrog, Cs. 2001. Hidrology and hidrogeology mapplotting – by kriging method. <http://lazarus.elte.hu/hun/digkonyv/bodrog/5.htm>, 01-09-2007.
- Geiger J. 2004. Geostatistika. <http://www.sci.u-szeged.hu/foldtan/Geostatistika.pdf>, 49 p., 03-09-2007.
- Korpás, A. 1996. General statistics, National Textbook Publisher, Budapest, 24-27 p.
- Meloni, J.C. 2003. PHP, MySQL and Apache usage. Panem Kft, Budapest.
- Oliver, M.A., Webster, R. 1990. "Kriging: a method of interpolation for geographical information system", *INT. J. Geographical Information Systems*, Vol. 4, No. 3, 313-332
- Soil Information Monitoring Professional Committee. 1995. Soil Information Monitoring System Methodology, Budapest
- Stein, A., Hoogerwerf, M., Bouma, J. 2003. *Geoderma* Vol. 43, Issues 1-2. 163-167
- Várallyai, Gy. 1993. Soil data bases for sustainable land use - Hungarian case study. In *Soil Resilience and Sustainable Land Use* (Eds.: Greenland, D. J. & Szabolcs, I.) 469-495. CAB International. Wallingford.
- Young, D.S. 1986. Indicator kriging for unit vectors: Rock joint orientations. Michigan Technological University, Houghton <http://www.springerlink.com/content/n038515710077387/?p=2337f7ffcb78402ea7e4459c0dcaec66&pi=0>, 27-02-2008.